Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-11 (2016) 562–569

Heavy-duty truck battery failure prognostics using random survival forests

Sergii Voronov, Daniel Jung, and Erik Frisk

Department of Electrical Engineering, Linköping University, Sweden
{sergii.voronov, daniel.jung, erik.frisk}@liu.se.

Abstract: Predicting lead-acid battery failure is important for heavy-duty trucks to avoid unplanned stops by the road. There are large amount of data from trucks in operation, however, data is not closely related to battery health which makes battery prognostic challenging. A new method for identifying important variables for battery failure prognosis using random survival forests is proposed. Important variables are identified and the results of the proposed method are compared to existing variable selection methods. This approach is applied to generate a prognosis model for lead-acid battery failure in trucks and the results are analyzed.

Keywords: Battery failure prognosis, Random survival forests, Variable selection

1. INTRODUCTION

Heavy-duty trucks are important for transporting goods, working at mines, or construction sites and it is vital that vehicles have a high degree of availability. In particular, this means to avoiding unplanned stops by the road which does not only cost due to the delay in delivery, but can also lead to damaged cargo.

One cause of unplanned stops is a failure in the electrical power system, and in particular the lead-acid starter battery. The main purpose of the battery is to power the starter motor to get the diesel engine running, but it is also used to, for example, power auxiliary units such as heating and kitchen equipment.

Prognostics and health management is an important part to prevent unexpected failures by more flexible maintenance planning. The purpose is to replace the battery before it fails but avoid changing it too often. Coarsely, there are two main approaches in prognostics, data-driven and model-based techniques but also hybrid approaches that combines the two are possible. Model-based prognostics utilizes a model of the monitored system and the fault to monitor to predict the degradation rate and Remaining Useful Life (RUL), see for example (Daigle and Goebel, 2011). Statistical data-driven methods generate a prediction model based on training data to predict RUL, see for example (Si et al., 2011), and is the approach followed here.

The main contribution in this work is a data-driven method to identify important variables from a set of variables, where many are not relevant for lead-acid battery failure prognosis, and use them to build prognostic models. The goal is to find important variables to design a battery failure prognostics model for automotive applications based on random survival forests (Ishwaran et al., 2008). This type of analysis is also important to better understand which factors that are correlated with battery failure rate and also what is causing it.

The outline is as follows. The problem is motivated in Section 2 and some background on random survival forests and variable importance are given in Section 3. Evaluation of existing methods for variable importance in random survival forests is presented in Section 4 showing the need for methodological developments in variables selection. The proposed variable selection method is described in Section 5. Then, the method is analyzed in detail in Section 6 and used to generate a random survival forest prognostic model in Section 7. Finally, some conclusions are presented in Section 8.

2. PROBLEM MOTIVATION

The prognostic problem studied here is to estimate a battery lifetime prediction function based on recorded vehicle data. The lifetime prediction function is defined as

$$\mathcal{B}^{\nu}(t;t_0) = P(T > t + t_0 \mid T \geq t_0,\ \nu)$$

where T is the random variable failure time of the battery and ν the vehicle data at $t = t_0$. The function $\mathcal{B}^{\nu}(t;t_0)$ is a function of t and gives the probability that the battery will function at least t time units after t_0. The data ν is recorded operational data for a specific vehicle which is further described in Section 2.1.

The reliability function (Cox and Oakes, 1984) is defined as

$$R(t) = P(T \geq t) \tag{1}$$

which is the probability that the battery of the specific vehicle will survive at least t time units. Then, the battery lifetime prediction function can be rewritten using the reliability function as

$$\mathcal{B}^{\nu}(t;t_0) = P(T > t + t_0 \mid T \geq t_0,\ \nu) = \frac{R^{\nu}(t+t_0)}{R^{\nu}(t_0)}. \tag{2}$$

Random Survival Forests (RSF) is a data-driven method that can be used for computing maximum-likelihood estimates of the reliability function, as illustrated by Fig. 1. The main objective in this work is to use Random Survival Forests to identify, from data, which vehicle data that is relevant for building RSF models to predict battery failures.

2.1 Operational data

In this work a vehicle fleet database is provided, where one snapshot of data is available from each vehicle including information regarding how the truck has been used and the configuration of the specific truck. There is also information if the battery has failed or not. The database contains a lot of information from the truck, not always related to battery

2405-8963
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.08.082

Fig. 1. A random survival forest computes the maximum likelihood estimate $\hat{R}^{\nu}(t)$ of the reliability function given a vehicle represented by the data ν. With the estimate $\hat{R}^{\nu}(t)$, the battery lifetime prediction function $\mathcal{B}^{\nu}(t;t_0)$ in (2) can be computed.

degradation, meaning that it is not known what available information is relevant for this specific task. Therefore, it is relevant to identify which variables are relevant for predicting battery lifetime. Previous works considering this vehicle data set are presented in (Frisk et al., 2014) and (Frisk and Krysander, 2015).

The choice of using RSF is motivated by the properties of the available database. Its main characteristics can be summarized as follows:

- 33603 vehicles from 5 EU markets
- 284 variables stored for each vehicle snapshot
- A single snapshot per vehicle
- Heterogeneous data, i.e., it is a mixture of categorical and numerical data
- Availability of histogram variables
- Censoring rate more than 90 percent
- Significant missing rate

The database contains different types of variables, including both categorical and numerical data. The censoring rate refers to that less than 10 percent of the vehicles in the database have had battery failures. This means that for most vehicles it is not known how long the battery will last. Also, there is a significant amount of missing data for the different vehicles, a property of database handled by RSF. One reason for the missing rate is due to the fact that data was recorded for different type of vehicles for which some variables are not applicable.

Another main characteristic of the database is that there are no time series available for a vehicle. It means that there is only one snapshot ν of the variables in the database for each vehicle. Information describing how the vehicle has been used is stored as histogram data where different variables represent how often specific sensor data is measured within different intervals. For example, there is a histogram describing how much time the vehicle has been subjected to different ambient temperatures.

When applying RSF to the data in the database, the objective is to find classes of vehicles with similar battery degradation properties. The reliability computed for a given class is an approximation of the true vehicle reliability which can be used to prognose battery failure. Due to the non-specific purpose of the database, it is probable that only small number of variables from set ν influence prediction of the battery failure rate. Thus, identifying the important variables in order to remove irrelevant ones, may improve the performance of a battery prognosis model. This problem is considered and explained in the successive subsection.

2.2 Variable selection using Random Survival Forests

The problem of identifying a set of important variables from a large set of variables is a relevant topic in machine learning, usually referred to as variable, or feature, selection, see (Guyon and Elisseeff, 2003). There are several reasons why variable selection is important when working with data-driven models. First, it is possible to improve the prediction performance by reducing the number of variables, for example, the quality of the predictor may become bad if the number of noisy variables (those that have no effect on battery failures) is large.

In the following illustrative example, two RSF are trained using synthetic data to show how the number of noisy variables can have a negative impact on prognostics performance.

Synthetic data is created with the following properties. Let h_0 be a constant nominal hazard rate h_0 for battery failure. The hazard rate

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid t \leq T)}{dt} \tag{3}$$

represents the probability of a battery failure at a particular time t, see (Cox and Oakes, 1984) for more details. In this example, the hazard rate does not change with time and the nominal hazard rate corresponds to an expected 10 years of battery life. It is assumed that there is one variable v_1 that explains how vehicle usage profile influences failure rate and changes h_0 to three hazard rates

$$h = \begin{cases} 1 \cdot h_0, & \text{if } v_1 = 1 \\ 2 \cdot h_0, & \text{if } v_1 = 2 \\ 3 \cdot h_0, & \text{if } v_1 = 3. \end{cases} \tag{4}$$

The scaling factors show how particular usage of the vehicle, described by v_1, changes the failure rate. Thus, there are three classes of batteries with different degradation profiles. Data for 3000 vehicles is generated with a censoring rate about 80 percent. The censoring rate is selected high to resemble the real vehicle database since censoring rate significantly affects the prediction performance of the RSF model. Two models with different numbers of noisy variables are considered to observe how it changes the RSF prediction.

In the first dataset, two noisy variables are added in addition to v_1, and in the second dataset, 100 noisy variable are added to v_1. After generating two RSF models, one for each dataset, one vehicle from each degradation profile is sampled from validation data and fed to the forest to generate predictions. It is shown in Fig. 2 (a) that predictions from the RSF for the case of 2 noisy variables (dashed blue curves) are following the theoretical reliability functions (red solid curves) significantly better than the predictions from the RSF for the case with 100 noisy variables, see Fig. 2 (b). Note that comparing the results shows a larger number of noisy variables results in worse prediction. The estimated reliability functions follow the theoretical values better with fewer noisy variables. This is something that can be expected.

One measure to evaluate prediction performance of RSF is error rate which should be low and is discussed further in Section 3. The error rate for the case with two noisy variables is 0.4088, for the case with 100 noisy variables is 0.4188. An important observation is that both cases give comparable error rates. However, Fig. 2 shows that there is a significant difference between the two predictors indicating the limitations of using error rate as a performance measure. The given situation happens due to the fact that for the case with a large number of noisy variables, it is hard for the model-building algorithm to find the relevant variables.

This example is illustrative, showing the effects of keeping a lot of noisy variables when generating the RSF model. The true reliability curves are in general unknown but the evaluation